Absolute Risk Estimation Using Nested Case-Control Data

Krystyn Villaflorces
Supervised by Dr. Agus Salim
La Trobe University

Contents

1 Introduction 3

2 Methods 5
   2.1 Survival Data ................................................................. 5
   2.2 Proportional Hazards Model ................................................. 5
   2.3 Cohort Design ................................................................. 6
   2.4 Nested Case-Control Design ............................................... 6
   2.5 Langholz-Borgan Approach vs. Weighted Approach ................. 6

3 Simulation 7
   3.1 Cohort Data ................................................................. 7
   3.2 Nested Case-Control Studies .............................................. 8
   3.3 No Matching ................................................................. 9
      3.3.1 Langholz-Borgan Approach ....................................... 9
      3.3.2 Weighted Likelihood Approach .................................. 9
3.4 Matching Based on Gender ........................................ 10
  3.4.1 Langholz-Borgan Approach .............................. 10
  3.4.2 Weighted Likelihood Approach ......................... 10

3.5 Matching Based on Gender and Age Category .............. 11
  3.5.1 Langholz-Borgan Approach .............................. 11
  3.5.2 Weighted Likelihood Approach ......................... 11

4 Results ....................................................................... 12
  4.1 No Matching ...................................................... 12
  4.2 Matching Based on Gender .................................... 13
  4.3 Matching Based on Gender and Age Category .......... 14

5 Discussion .................................................................... 16

6 Acknowledgements ..................................................... 18

7 References .................................................................... 18
1 Introduction

There has recently been a marked increase in the development of prognostic models to predict the risk associated with certain diseases. The main goal of a disease prediction model is estimating absolute risk, which is essential as it is the basis of treatment recommendations and lifestyle interventions. The importance of estimating this risk precisely cannot be understated, as only those whose risk is estimated to be over a particular threshold will be recommended treatment. Ill effects of overestimation and underestimation are evident, with individuals consequently being prescribed unnecessary treatments or foregoing preventative measures, respectively.

It is standard for researchers who develop risk prediction models to use a cohort study design as it gives rise to valid absolute risk estimates. Though this is the optimal design in terms precision and statistical efficiency, budgetary and time constraints can mean this costly method is often impractical. Accordingly, a trend has emerged in using only a subset of subjects from large cohorts to counter this. One such method for using only a particular subcohort, that is the focus of this report, is a Nested Case-Control (NCC) design. An NCC design involves matching a predetermined number of controls to cases, such that each case is included in the subcohort. More specifically, at each time of onset, a number of individuals are randomly chosen from the risk set, that is, those free from disease at that onset time, to match to the case. Cai and Zheng (2012) showed that using only 3 controls per case resulted in a minimal loss of statistical efficiency. Often under this design, as a means of reducing variability and hence increasing power, cases are matched to controls on various characteristics such as age or gender. While using a matched design can be preferable, it has been shown to cause complications in obtaining precise or unbiased estimates of risk when matching variables are also risk factors (Ganna et al., 2012).

When conducting a study that uses only a subset of individuals from a large cohort, the probability of being selected is not the same for all individuals, rather it depends on their status with regards to the disease being studied and, when using a matched design, on their matching variables. Hence, when analysing the subcohort data, adjustments need to be made to take into account the now increased prevalence by factoring in these different probabilities. Furthermore, estimation is increasingly complicated when a matched design is desired, as the number of eligible controls at each onset time is reduced with the addition of each matching variable. Ganna et al. (2012) mentions the complexity of calculating
appropriate weights in such a design, and states that coefficients for matching variables are inestimable, leading to biased calculations of individual risk. This finding was a strong motivating factor, giving purpose to this project to show that this is not in fact the case. That is, it was our aim to demonstrate that unbiased, accurate estimates can be obtained by employing the correct statistical methodology. We proposed a weighted likelihood method for analysing NCC data, combining inverse probability weighting and a method known as “breaking the matching” (Salim et al., 2009; Samuelsen, 1997). The way in which we set out to test this was to compare the effectiveness of our method to an existing approach developed by Langholz and Borgan (1997) in estimating absolute risk from simulated NCC data under different matching designs.
2 Methods

As previously stated, it was our aim to compare our proposed weighted likelihood method to the Langholz-Borgan approach by way of simulation studies. Before detailing the process in which this was done, some additional information on background theory and methodology needs to be introduced.

2.1 Survival Data

When dealing with data from medical and epidemiological studies, one needs to treat it carefully as it is often characterised as survival data. This is data in the form of times from a well-defined time origin up until some time of onset, failure or death. The reason in which standard methods are inappropriate is the presence of censored survival times. Censoring for an individual is most commonly seen when the end-point of interest has not occurred for that individual, often due to loss from follow-up, death from an unrelated cause, or simply still being alive at the completion of a study (Collett, 2003). Analysis of survival data focuses on the risk, or hazard, of death at any time after commencing the study and as a result hazard is modelled directly.

2.2 Proportional Hazards Model

One of the most commonly utilised models for analysing survival data is the Cox Proportional Hazards (PH) model. This describes the relation between the hazard and a set of covariates of interest. The hazard for the $i^{th}$ individual under this model is

$$
\lambda_i(t) = \lambda_0(t) e^{\beta x_i + \gamma z_i},
$$

where $\lambda_0(t)$ is the baseline hazard, $x_i$ is the vector of variables of interest and $z_i$ is the vector of other variables that both affect risk and can be used as matching variables.

In order to estimate the unknown $\beta$ and $\gamma$ coefficients we use the values that maximise the Cox Partial Likelihood, given by

$$
L(\beta, \gamma) = \prod_{i=1}^{r} \frac{e^{x_i^T \beta + z_i^T \gamma}}{\sum_{j \in R(t_i)} e^{x_j^T \beta + z_j^T \gamma}},
$$

where $r$ is the number of events, the product is over all cases and the summation is over the individuals in the risk set at time $t_i$.

Commonly, researchers are only interested in determining the particular risk factors associated with a disease of interest, and so obtaining estimates for $\beta$ and $\gamma$ is all that is required. Since our aim was to estimate absolute risk, we additionally needed to obtain estimates for the cumulative baseline hazards, found by using the Breslow estimator

$$
\Lambda_0(t) = \sum_{i=1}^{r} \frac{I(t_i \leq t) d_i}{\sum_{j \in R(t_i)} e^{x_j^T \beta + z_j^T \gamma}},
$$
where $d_i$ is the number of events and $R(t_i)$ is the risk set at the $i^{th}$ event time $t_i$.

Once these estimates have been obtained, the absolute risk for the $i^{th}$ individual can be calculated as

$$F_i(t) = 1 - \exp(-\Lambda_0(t))\exp(x_i\beta + z_i\gamma) \quad (4)$$

### 2.3 Cohort Design

Data collected from large cohorts is seen as the gold standard as it allows estimates to be made that are both precise and unbiased. It is commonly the case that such cohorts consist of tens of thousands of individuals who are followed for several years. These individuals all start healthy at baseline and over the course of the study a multitude of measurements are taken. Although precise estimates can somewhat easily be attained, it is often an impractical design since there is generally a large associated cost with obtaining and analysing such large data sets. This is a result of it being very expensive to measure certain things, especially when genes and biomarkers are involved.

### 2.4 Nested Case-Control Design

In an effort to control the costs involved with a full cohort study, different study designs have arisen that involve only using a subset of participants, such as case-control, case-cohort and nested case-control. We focus on the Nested Case-Control (NCC) design in this paper. As previously described, a given number of controls are compared to the individuals who experience onset or death, with all cases being analysed. At each time point in which an onset occurs, currently healthy individuals are matched for comparison, continuing until each case has controls matched to it. In this fashion, individuals can be selected more than once as controls, or those chosen as controls at one time point may later turn into cases. This is another complication that is dealt with by our proposed weighted approach.

### 2.5 Langholz-Borgan Approach vs. Weighted Approach

All of the formulae given in the previous section are those that are utilised for calculations associated with full cohort designs. Adjustments need to be made to these in order to analyse the subcohorts, and this is done differently under the two methods that we address in this report. While the L-B approach uses weighting based on a time-dependent ratio of the number of eligible controls to the number of controls chosen to be matched to the cases, the weighted likelihood approach calculates individual weights based on a function of the probability of inclusion into, taking into account all possible onset times. When matching is present, these give very different results which are given, as well as the way in which they are utilised within the different formulae, in the following sections.
3 Simulation

3.1 Cohort Data

We performed simulations by generating data assumed to follow the given PH model, where the hazard for the \( i \)th individual is given by (1), with the baseline hazard assumed to be constant. The values in this simulation were based on gender and age (variables in \( z_i \)), cholesterol, HDL, SBP, smoking status and antihypertensive treatment status (variables in \( x_i \)) for 50,000 cohort members. These values were generated using a multivariate normal distribution based on the following means and correlations between variables, as observed from an NCC study of coronary heart disease conducted within the Singapore Chinese Health Study (SCHS) (Hankin et al., 2001).

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Cholesterol</th>
<th>HDL</th>
<th>SBP</th>
<th>Antihypertensive Treatment Status</th>
<th>Smoking Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>63.128</td>
<td>0.513</td>
<td>202.903</td>
<td>53.765</td>
<td>135.557</td>
<td>0.270</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Cholesterol</th>
<th>HDL</th>
<th>SBP</th>
<th>Antihypertensive Treatment Status</th>
<th>Smoking Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>59.346</td>
<td>-0.584</td>
<td>13.418</td>
<td>14.390</td>
<td>45.395</td>
<td>0.467</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.584</td>
<td>0.229</td>
<td>-3.438</td>
<td>-1.905</td>
<td>0.017</td>
<td>-0.002</td>
</tr>
<tr>
<td>Chol</td>
<td>13.418</td>
<td>-3.438</td>
<td>1277.726</td>
<td>162.776</td>
<td>81.593</td>
<td>-1.612</td>
</tr>
<tr>
<td>HDL</td>
<td>14.390</td>
<td>-1.905</td>
<td>162.776</td>
<td>159.117</td>
<td>-14.988</td>
<td>-0.928</td>
</tr>
<tr>
<td>SBP</td>
<td>45.395</td>
<td>0.017</td>
<td>81.593</td>
<td>-14.988</td>
<td>487.537</td>
<td>1.928</td>
</tr>
<tr>
<td>Antihyp</td>
<td>0.467</td>
<td>-0.002</td>
<td>-1.612</td>
<td>-0.928</td>
<td>1.928</td>
<td>0.218</td>
</tr>
<tr>
<td>Smoke</td>
<td>-0.188</td>
<td>0.057</td>
<td>-0.406</td>
<td>-0.715</td>
<td>0.377</td>
<td>-0.026</td>
</tr>
</tbody>
</table>

Age values were rounded to the nearest integer and for the variables gender, smoking status and antihypertensive treatment status, the simulated values were further categorized into binary values, so that the observed means were equal to the population means from the SCHS. In practice, for each variable, this categorization is done by splitting the simulated
values into two using the $100m^{th}$ percentile value as cut-off point, where $m$ is the SCHS population mean of the variable. To obtain the appropriate percentile, we used the quantile function in the statistical package $R$.

Given the simulated values of the risk factors above, we generated the vector of time of onset for the disease using the inverse CDF method (Westfall and Henning, 2013), assuming the given PH model (with the following parameters from SCHS data: $\lambda_0 = 8.87 \times 10^{-6}, \gamma_{\text{Age}} = 0.05, \gamma_{\text{Gender}} = 0.47, \beta_{\text{Chol}} = 0.01, \beta_{\text{HDL}} = -0.02, \beta_{\text{SBP}} = 0.01, \beta_{\text{Antihyp}} = 0.29, \beta_{\text{Smoke}} = 0.54$) with a constant baseline hazard,

$$t = \frac{-\log(1-u)}{\lambda_0(t)e^{X \beta}},$$

(5)

where $u$ is a vector of randomly generated values from a uniform distribution, $X$ is the matrix of risk factor values (one row per subject) and $\beta$ is a given vector of true log hazard ratio values, again obtained from SCHS population. The constant baseline hazard was set to $\lambda_0(t) = 8.866098 \times 10^{-6}$. To simplify, we assumed that all individuals enter the study at time $t = 0$ so the time of onset is measured as time (years) since the start of study.

Using this set up we generated 500 cohorts, each of size $N = 50,000$, with random censoring time generated as an exponential variable with a rate of 0.1, giving an average length of follow-up of 10 years. Individuals with a censoring time occurring before the time of onset are censored, and a maximum censoring time was set at $t = 12$ years.

### 3.2 Nested Case-Control Studies

Within each of the 500 cohorts we generated three types of NCC studies where 2 randomly chosen controls were matched with each case. The 3 types were: NCC where cases and controls were not matched on any risk factors (no matching); NCC where controls needed to be of the same gender as the case (matching based on gender); and NCC where controls needed to be of the same gender and within the same age group as the case (matching based on gender and age category). For the third type of study, age was split into 4 categories of similar sizes using quartiles as cut-offs. For each NCC study, estimates of log hazard ratios, cumulative baseline hazards and absolute risk were calculated using both the Langholz-Borgan and weighted approaches, as well as cohort estimates that serve as the gold standard.
3.3 No Matching

3.3.1 Langholz-Borgan Approach

Log hazard ratio estimates were obtained using the full Cox PH model with all covariates and the data stratified by case-control triplets. These estimates were obtained by maximising

$$L(\beta, \gamma) = \prod_i \frac{e^{x_i \beta + z_i \gamma}}{\sum_{j \in R_i} e^{x_j \beta + z_j \gamma}}$$

(6)

where the product takes place over all cases and \(R_i\) is the selected risk set for the case at time \(t_i\). Given the log hazard ratio estimates, cumulative baseline hazards were estimated as

$$\Lambda_0(t) = \sum_{i=1}^{n_c} \frac{I(t_i \leq t)}{\sum_{j \in R_i} w(t_i) e^{x_j \beta + z_j \gamma}};$$

(7)

where \(w(t_i)\) is the weight prescribed by Langholz and Borgan (1997) given by:

$$\frac{M_i}{m};$$

(8)

where \(M_i\) is the size of the risk set at time \(t_i\), and \(m\) is the required number of controls per case.

The absolute risk for the \(i^{th}\) individual is calculated as

$$F_i(t) = 1 - \exp(-\Lambda_0(t)) \exp(x_i \beta + z_i \gamma)$$

(9)

3.3.2 Weighted Likelihood Approach

Any duplications of individuals selected more than once as controls were removed and the data was treated as a pool (i.e. rows of data are unique), which essentially breaks the matching. Using only these unique individuals, the Cox PH model was used to obtain the estimates of log hazard ratios by maximising

$$L(\beta, \gamma) = \prod_i \frac{e^{x_i \beta + z_i \gamma}}{\sum_{j \in R_i} w_j e^{x_j \beta + z_j \gamma}}$$

(10)

Given the log hazard ratio estimates, cumulative baseline hazards were estimated as

$$\Lambda_0(t) = \sum_{i=1}^{n_c} \frac{I(t_i \leq t)}{\sum_{j \in R_i} w_j e^{x_j \beta + z_j \gamma}};$$

(11)
where the summation is over all cases, \( R \) is the set of all selected cases and controls who were at risk at time \( t_i \) (time of onset of case \( i \)). The weights, \( w_j \), were calculated as the inverse of the probability of inclusion for individual \( j \), and the probability of inclusion for cases was set to 1 as all cases were selected. The probability of inclusion for controls is \( 1/p_j \), where \( p_j \) is calculated using

\[
1 - p_j = \prod_{i \in \Omega, s_j \leq t_i \leq t_j} \left[ 1 - \frac{m}{M_i} \right],
\]

(12)

where the product takes place at all onset times when individual \( j \) is eligible to be selected as a control, \( m \) and \( M_i \) are as defined before.

### 3.4 Matching Based on Gender

#### 3.4.1 Langholz-Borgan Approach

Log hazard ratio estimates were calculated using a Cox PH model with all covariates except for gender. These estimates were obtained by maximising (6), with \( \gamma_{Gender} \) no longer estimable as it will cancel out in the equation since the cases and controls have the same \( z \) value. To balance this, we allowed a different baseline hazard for each gender when estimating the PH model. This was done by adding gender as stratifying variable (in addition to the case-control triplets ID) when executing the \( R \) command \texttt{coxph}

\[
L(\beta, \gamma) = \prod_i e^{z_i \gamma_{Age}} \sum_{j \in R_i} \frac{e^{x_j \beta + x_j \gamma_{Age}}}{w(t_i) e^{x_j \beta + x_j \gamma_{Age}}}
\]

(13)

Cumulative baseline hazards were estimated for each gender group separately given the log hazard ratio estimates. The cumulative baseline hazard estimates for gender = G is given by

\[
\Lambda_{0,G}(t) = \sum_{i=1}^{n_c} \frac{I(t_i \leq t, Gender_i = G)}{\sum_{j \in R_i} w(t_i) e^{x_j \beta + x_j \gamma_{Age}}}
\]

(14)

Absolute risk was then calculated in the same manner as before, using (9), with the different cumulative baseline hazards being used, dependent on the gender of the individual and setting the gender coefficient \( \gamma_{Gender} \) to zero.

#### 3.4.2 Weighted Likelihood Approach

Again, any duplicates of controls selected more than once were removed and the individuals were pooled. Using the unique data, the Cox PH model was used to obtain the estimates of log hazard ratios using (10). Given those estimates, the cumulative baseline hazards were, again, estimated by (11), and absolute risk by (9).
3.5 Matching Based on Gender and Age Category

3.5.1 Langholz-Borgan Approach

In order to have enough controls to match to each of the cases, we created a new variable, *Age Category*, for matching rather than using exact age. Commonly this is done by classifying individuals into blocks consisting of a 5 or 10 year age range. However, due to the wide range of ages in the data, and time constraints, I chose to do this by creating 4 levels within the variable corresponding to the approximate quartiles, in order to divide the data into equally sized groups to ensure the existence of controls that matched the cases. As a result, the original variable, *Age*, was not technically confounded, and so I tested the L-B approach using models both with and without *Age* as a factor.

**Model without Age**

Coefficient estimates were obtained using the Cox PH model with all covariates except for age and gender, with the data stratified by gender, age category and case-control triplets. These estimates were obtained by maximising (6), however, we see that with matching on age and gender cases and controls have the same $z$ values, and so $\gamma$ cancels out, giving

$$L(\beta) = \prod_i \frac{e^{x_i\beta}}{\sum_{j \in R_i} e^{x_j\beta}}$$  \hspace{1cm} (15)

Cumulative baseline hazards were obtained for each age group/gender combination, again without $\gamma$, as

$$\Lambda_{0,A,G}(t) = \sum_{i=1}^{n_i} \frac{I(t_i \leq t, Age_i = A, Gender_i = G)}{\sum_{j \in R_i} w(t_i) e^{x_j\beta}}$$  \hspace{1cm} (16)

where $w(t_i)$ was as given before.

**Model with Age**

Coefficient estimates were found using the Cox PH model with all covariates except for gender, and the data stratified by gender, age category and case-control triplets to compensate, as before. The estimates were obtained again by maximising (13), again with $\gamma_{Gender}$ inestimable. Cumulative baseline hazards were acquired for each age group/gender combination separately, using the weighted Breslow estimator given by (14). Absolute risk was estimated using (9), again by substituting the correct baseline hazard estimates, dependent on the age and gender of the individual, and setting the $\gamma_{Gender}$ to zero.

3.5.2 Weighted Likelihood Approach

The method for simulating the weighted likelihood approach is the same for all of the matching designs, because in each type only unique rows of data are kept and then all treated as a pool.
4 Results

4.1 No Matching

To serve as a method for comparison, estimates were obtained for both approaches and for the full cohort where no variable matching for controls was used. Unbiased estimates for all methods were obtained and the results for each are as follows.

Table 3: Estimated Coefficients using L-B and Weighted Approaches with No Matching

<table>
<thead>
<tr>
<th></th>
<th>L-B Approach</th>
<th>Weighted Approach</th>
<th>Cohort</th>
<th>True Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est</td>
<td>SE</td>
<td>Est</td>
<td>SE</td>
</tr>
<tr>
<td>Age</td>
<td>0.0545</td>
<td>0.0061</td>
<td>0.0548</td>
<td>0.0062</td>
</tr>
<tr>
<td>Gender</td>
<td>0.4682</td>
<td>0.0866</td>
<td>0.4685</td>
<td>0.0864</td>
</tr>
<tr>
<td>Chol</td>
<td>0.0057</td>
<td>0.0013</td>
<td>0.0057</td>
<td>0.0013</td>
</tr>
<tr>
<td>HDL</td>
<td>-0.0230</td>
<td>0.0038</td>
<td>-0.0231</td>
<td>0.0038</td>
</tr>
<tr>
<td>SBP</td>
<td>0.0130</td>
<td>0.0020</td>
<td>0.0130</td>
<td>0.0020</td>
</tr>
<tr>
<td>Antihyp</td>
<td>0.2850</td>
<td>0.0978</td>
<td>0.2859</td>
<td>0.0978</td>
</tr>
<tr>
<td>Smoke</td>
<td>0.5378</td>
<td>0.1034</td>
<td>0.5400</td>
<td>0.1042</td>
</tr>
</tbody>
</table>

From Table 3, we can see that the L-B and weighted estimates are very close to or equal to the full cohort estimates and true values. It can be seen that the cohort estimates have achieved the most precision, with standard errors noticeably lower than that of the other two methods.

Figure 1: Absolute risk estimates for NCC vs Full Cohort with no matching for one typical realisation using: (a) L-B approach, and (b) Weighted approach.
Estimation of absolute risk was also precise and unbiased, with the mean individual risk under the L-B approach calculated at 0.0286, in comparison to 0.0282 for the full cohort. Estimation of absolute risk using the weighted approach was slightly more accurate than those of the L-B approach, with a mean individual risk of 0.0283. This is illustrated in Figure 1. To get an appreciation for the spread of the estimates, the mean square error (taken to be the average squared distance between the cohort and each of the methods) was calculated as $9.2980 \times 10^{-6}$ and $9.3246 \times 10^{-6}$ for the L-B and weighted approaches, respectively.

4.2 Matching Based on Gender

As an intermediate test, we compared the estimates obtained by both methods to the full cohort and known true values, when using a matched design based on a single variable with a binary response, Gender.

Table 4: Estimated Coefficients using L-B and Weighted Approaches with Matching Based on Gender

<table>
<thead>
<tr>
<th></th>
<th>L-B Approach</th>
<th>Weighted Approach</th>
<th>Cohort</th>
<th>True Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est</td>
<td>SE</td>
<td>Est</td>
<td>SE</td>
</tr>
<tr>
<td>$\gamma_{Age}$</td>
<td>0.0541</td>
<td>0.0063</td>
<td>0.0542</td>
<td>0.0062</td>
</tr>
<tr>
<td>$\gamma_{Gender}$</td>
<td>0</td>
<td>0</td>
<td>0.4663</td>
<td>0.0802</td>
</tr>
<tr>
<td>$\beta_{Chol}$</td>
<td>0.0058</td>
<td>0.0013</td>
<td>0.0058</td>
<td>0.0013</td>
</tr>
<tr>
<td>$\beta_{HDL}$</td>
<td>-0.0230</td>
<td>0.0038</td>
<td>-0.0231</td>
<td>0.0038</td>
</tr>
<tr>
<td>$\beta_{SBP}$</td>
<td>0.0131</td>
<td>0.0020</td>
<td>0.0132</td>
<td>0.0020</td>
</tr>
<tr>
<td>$\beta_{Antihyp}$</td>
<td>0.2815</td>
<td>0.1003</td>
<td>0.2800</td>
<td>0.0988</td>
</tr>
<tr>
<td>$\beta_{Smoke}$</td>
<td>0.5428</td>
<td>0.0907</td>
<td>0.5427</td>
<td>0.0928</td>
</tr>
</tbody>
</table>

We can see again in Table 4 that accurate estimates were obtained, however, as previously mentioned $\gamma_{Gender}$ is not estimable under the L-B approach as it has been used as a matching variable. This is not an issue in the weighted approach and a close estimate is obtained. Both methods perform with similar accuracy and variability and also when compared to estimates obtained from the full cohort, though the cohort estimates still have a smaller SE.

The average individual risks were almost the same, calculated as 0.0283, 0.0282 and 0.0282 for the L-B approach, weighted approach and full cohort, respectively. The mean squared error for L-B was $8.0398 \times 10^{-6}$, slightly higher than that of the weighted approach, $7.8255 \times 10^{-6}$. 

Postcard Address: 111 Barry Street
C/- The University of Melbourne
Victoria 3010 Australia

Email: enquiries@amsi.org.au
Phone: +61 3 8344 1777
Fax: +61 3 9349 4106
Web: www.amsi.org.au

13
4.3 Matching Based on Gender and Age Category

To test the feasibility of both methods under a more complicated matching design, we ran a simulation using both binary and non-binary matching variables, \texttt{Gender} and \texttt{Age Category}, respectively. Estimates were again compared to the full cohort and known true values to gauge precision.

We can see from Table 5 that the L-B approach not using age as a part of the Cox PH model consistently had the least accurate estimates, as well as not being able to estimate $\gamma_{\text{Age}}$ or $\gamma_{\text{Gender}}$. The model that did include age for the L-B approach performed similarly to the weighted approach and the full cohort, though $\gamma_{\text{Gender}}$ was inestimable.

Table 5: Estimated Coefficients using L-B and Weighted Approaches with Matching Based on Gender and Age Category

<table>
<thead>
<tr>
<th></th>
<th>L-B Model w/o Age</th>
<th>L-B Model w/ Age</th>
<th>Weighted</th>
<th>Cohort</th>
<th>True Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est</td>
<td>SE</td>
<td>Est</td>
<td>SE</td>
<td>Est</td>
</tr>
<tr>
<td>$\gamma_{\text{Age}}$</td>
<td>0</td>
<td>0</td>
<td>0.0546</td>
<td>0.0139</td>
<td>0.0548</td>
</tr>
<tr>
<td>$\gamma_{\text{Gender}}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.4697</td>
</tr>
<tr>
<td>$\beta_{\text{Chol}}$</td>
<td>0.0056</td>
<td>0.0013</td>
<td>0.0058</td>
<td>0.0013</td>
<td>0.0058</td>
</tr>
<tr>
<td>$\beta_{\text{HDL}}$</td>
<td>-0.0226</td>
<td>0.0036</td>
<td>-0.0236</td>
<td>0.0036</td>
<td>-0.0236</td>
</tr>
<tr>
<td>$\beta_{\text{SBP}}$</td>
<td>0.0140</td>
<td>0.0020</td>
<td>0.0131</td>
<td>0.0021</td>
<td>0.0131</td>
</tr>
<tr>
<td>$\beta_{\text{Antihyp}}$</td>
<td>0.2741</td>
<td>0.0931</td>
<td>0.2859</td>
<td>0.0933</td>
<td>0.2848</td>
</tr>
<tr>
<td>$\beta_{\text{Smoke}}$</td>
<td>0.5451</td>
<td>0.0953</td>
<td>0.5433</td>
<td>0.0962</td>
<td>0.5431</td>
</tr>
</tbody>
</table>
Although, again very similar mean individual risks were obtained for each method, given as 0.0283, 0.0283, 0.0281 and 0.0281, for L-B without age, L-B with age, Weighted and Cohort, respectively, the mean squared error showed much more notable differences. The MSE ratio between L-B without age and L-B with age was 3.57, given by the individual MSE’s of $5.6295 \times 10^{-5}$ and $1.5768 \times 10^{-5}$. The MSE ratio between L-B without age and weighted was 9.79, with an MSE of $5.7470 \times 10^{-6}$. The difference in the spread of risk estimates can be seen in Figures 3 and 4.

![Figure 3: Absolute risk estimates for NCC vs Full Cohort with matching based on gender and age category for one typical realisation using the L-B approach for: (a) model without age, and (b) model with age.](image)

![Figure 4: Absolute risk estimates for NCC vs Full Cohort for one typical realisation using the weighted approach with matching based on gender and age category.](image)
5 Discussion

This paper has shown the feasibility of using our proposed weighted approach in obtaining precise, unbiased estimates of risk in a matched NCC study. This was done in an effort to use all covariates, by breaking the matching, where each unique contribution of an individual to the log-likelihood was weighted by the inverse of the probability of the individual being included in the study. This was demonstrated by simulating realistic cohorts, from which NCC data was obtained, and then comparing its efficiency both to the full cohort and to a widely accepted existing method put forth by Langholz and Borgan (1997).

It was shown that the Langholz-Borgan approach was effective in obtaining precise estimates in an unmatched NCC design, confirming previous findings, however, estimation became problematic when matching was present. Problems were less pronounced when matching was done with a single binary variable, but as the number and complexity of the matching variables were increased, so too was the lack of precision. The weighted likelihood approach worked comparably well in an unmatched design, but very evidently outperformed the Langholz-Borgan approach under the matched designs, shown clearly by the distance measures and the plots of risk estimates.

One of the main differences between the two approaches, and consequently the difference in results, is in the method of weighting used for calculating the estimates. The purpose of using inverse probability weighting was to control for the fact that prevalence was higher in the NCC data than the full cohort, and the risk sets at each time point had changed, due to matching. The Langholz-Borgan approach calculates weights at each onset time, based on the ratio of eligible controls to selected controls, which consequently means that weighting is a function of time. This means that individuals selected as controls more than once will be given different weights at the different times they are selected, depending on the size of the risk set at each of those time points. In contrast, the weighting used in our proposed weighted likelihood method is calculated for each individual, done across all possible onset times, meaning that it is not a function of time.

Another of the differences is that our weighted likelihood method “breaks” the matching, treating only a unique set of pooled data. This meant that no adjustments were needed between the different designs when the simulations were conducted using this method. Interestingly, although the matching was broken, we still see residual effects in that there is a reduction in the standard error of the estimates from the unmatched to the gender matched to the age category and gender matched design. This further illustrates the appropriateness of this approach over Langholz and Borgan’s when a matched design is desired.
The fact that the simulations were done using realistic parameters that were estimated from a large cohort study means that we can be more confident making inferences about the results than if we had generated random data. The use of different matching designs utilising different numbers of and different types of matching variables gives us more of an insight into the nature of the relationship between the two approaches and the covariates. However, our findings would be strengthened by further collaborating evidence that utilised real datasets or made comparisons between NCC designs, case-control and case-cohort designs.

In conducting our simulations over different scenarios with the two approaches we achieved our project aim in obtaining findings that refute the result stated in Ganna et al. (2012). If the correct, rigorous methodology is applied, valid and unbiased risk estimates are achievable. With more studies in support of our findings, it will be seen that utilising our methods are preferable over others, given appropriate conditions. This is a very important finding relevant to several fields, such as biological or epidemiological studies in which there is difficulty or large expense associated with covariate measurement, or for diseases associated with gender, age or other factors, that make a matched design a necessity. This certainly paves a new way for researchers to significantly reduce time, effort and costs whilst compromising very little statistical efficiency.
6 Acknowledgements

I would like to thank my supervisor, Dr. Agus Salim, for providing me with a fantastic project, and for his immense support throughout. I would also like to thank AMSI for funding this project and giving me the opportunity to get a taste for research, and CSIRO for organising the Big Day In, giving us the chance to present findings, network and learn about the research undertaken by other vacation scholars.

7 References


Salim, A. et al. (2009). “Combining data from 2 nested case–control studies of Combining data from 2 nested case–control studies of overlapping cohorts to improve efficiency”. In: Biostatistics 10.1, pp. 70–79.


Krystyn Villaflores received a 2013/14 AMSI Vacation Research Scholarship